Connecting Vision and Language with Localized Narratives

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo,
Radu Soricut, and Vittorio Ferrari

Google Research

Abstract. We propose Localized Narratives, a new form of multimodal image annotations connecting vision and language. We ask annotators to describe an image with their voice while simultaneously hovering their mouse over the region they are describing. Since the voice and the mouse pointer are synchronized, we can localize every single word in the description. This dense visual grounding takes the form of a mouse trace segment per word and is unique to our data. We annotated 849k images with Localized Narratives: the whole COCO, Flickr30k, and ADE20K datasets, and 671k images of Open Images, all of which we make publicly available. We provide an extensive analysis of these annotations showing they are diverse, accurate, and efficient to produce. We also demonstrate their utility on the application of controlled image captioning.

1 Introduction

Much of our language is rooted in the visual world around us. A popular way to study this connection is through Image Captioning, which uses datasets where images are paired with human-authored textual captions [8, 59, 46]. Yet, many researchers want deeper visual grounding which links specific words in the caption to specific regions in the image [32, 33, 44, 45]. Hence Flickr30k Entities [40] enhanced Flickr30k [59] by connecting the nouns from the captions to bounding boxes in the images. But these connections are still sparse and important aspects remain ungrounded, such as words capturing relations between nouns (as "holding" in "a woman holding a balloon"). Visual Genome [27] provides short descriptions of regions, thus words are not individually grounded either.

In this paper we propose Localized Narratives, a new form of multimodal image annotations in which we ask annotators to describe an image with their voice while simultaneously hovering their mouse over the region they are describing. Figure 1 illustrates the process: the annotator says "woman" while using their mouse to indicate her spatial extent, thus providing visual grounding for this noun. Later they move the mouse from the woman to the balloon following its string, saying "holding". This provides direct visual grounding of this relation. They also describe attributes like "clear blue sky" and "light blue jeans". Since voice is *synchronized* to the mouse pointer, we can determine the image location of every single word in the description. This provides dense visual grounding in the form of a mouse trace segment for each word, which is unique to our data.

In order to obtain written-word grounding, we additionally need to transcribe the voice stream. We observe that automatic speech recognition [15, 1, 41] typically results in imperfect transcriptions. To get data of the highest quality, we

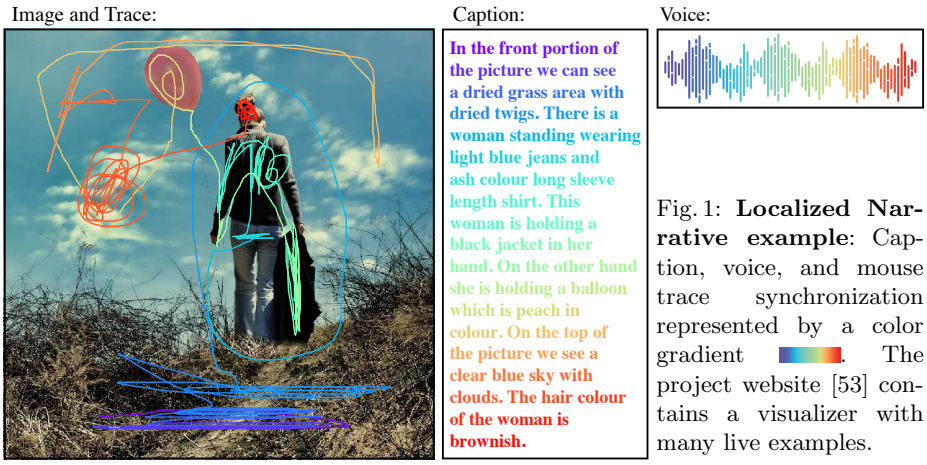


Fig. 1: **Localized Narrative example**: Caption, voice, and mouse trace synchronization represented by a color gradient . The project website [53] contains a visualizer with many live examples.

ask annotators to transcribe their own speech, immediately after describing the image. This delivers an accurate transcription, but without temporal synchronization between the mouse trace and the written words. To address this issue, we perform a sequence-to-sequence alignment between automatic and manual transcriptions, which leads to accurate *and* temporally synchronized captions. Overall, our annotation process tightly connects four modalities: the image, its spoken description, its textual description, and the mouse trace. Together, they provide dense grounding between language and vision.

Localized Narratives is an efficient annotation protocol. Speaking and pointing to describe things comes naturally to humans [22, 38]. Hence this step takes little time (40.4 sec. on average). The manual transcription step takes 104.3 sec., for a total of 144.7 sec. This is lower than the cost of related grounded captioning datasets Flickr30k Entities and Visual Genome [27, 40], which were made by more complicated annotation processes and involved manually drawing bounding boxes (Sec. 4.1 – Annotation Cost). Moreover, if automatic speech recognition improves in the future it might be possible to skip the manual transcription step, making our approach even more efficient.

We collected Localized Narratives at scale: we annotated the whole COCO [31] (123k images), ADE20K [62] (20k) and Flickr30k [59] (32k) datasets, as well as 671k images of Open Images [29]. We make the Localized Narratives for these 848,749 images publicly available [53]. We provide an extensive analysis (Sec. 4) and show that: (i) Our data is rich: we ground all types of words (nouns, verbs, prepositions, etc.), and our captions are substantially longer than in most previous datasets [8, 59, 27, 46]. (ii) Our annotations are diverse both in the language modality (e.g. caption length varies widely with the content of the image) and in the visual domain (different pointing styles and ways of grounding relationships). (iii) Our data is of high quality: the mouse traces match well the location of the objects, the words in the captions are semantically correct, and the manual transcription is accurate. (iv) Our annotation protocol is more efficient than for related grounded captioning datasets [27, 40].

Image	Text	Speech	Grounding	Task
In	Out	-	-	Image captioning [51, 55, 52], Paragraph generation [26, 56, 64]
Out	In	-	-	Text-to-image Generation [42, 47, 57]
In	Out	-	Out	Dense image captioning [21, 58, 25], Dense relational captioning [25]
In	Out	-	In	Controllable and Grounded Captioning [10]
In	In	-	Out	Phrase grounding [13]
In	In + Out	-	-	Visual Question Answering [4, 34, 20]
In	In + Out	-	Out	Referring Expression Recognition [24, 35, 9]
In	-	In	Out	Discover visual objects and spoken words from raw sensory input [18]
-	In	Out	-	Speech recognition [15, 1, 41]
-	Out	In	-	Speech synthesis [37, 23, 36]
Out	In	-	In	Image generation/retrieval from traces
In	Out	In	In	Grounded speech recognition
In	-	In	Out	Voice-driven environment navigation

Table 1: **Tasks enabled by Localized Narratives**. Each row represents different uses of the four elements in a Localized Narrative: image, textual caption, speech, and grounding (mouse trace); labeled as being input (In) or output (Out) for each task.

Since Localized Narratives provides four synchronized modalities, it enables many applications (Tab. 1). We envision that having each word in the captions grounded, beyond the sparse set of boxes of previous datasets [40, 24, 35, 27], will enable richer results in many of these tasks and open new doors for tasks and research directions that would not be possible with previously existing annotations. As a first example, we show how to use the mouse trace as a fine-grained control signal for a user to request a caption on a particular image (Sec. 5). Mouse traces are a more natural way for humans to provide a sequence of grounding locations, compared to drawing a list of bounding boxes [10]. We therefore envision its use as assistive technology for people with imperfect vision. In future work, the mouse trace in our Localized Narratives could be used as additional attention supervision at training time, replacing or complementing the self-supervised attention mechanisms typical of modern systems [3, 46, 7, 60, 55]. This might train better systems and improve captioning performance at test time, when only the image is given as input. Alternatively, our mouse traces could be used at test time only, to inspect whether current spatial attention models activate on the same image regions that humans associate with each word.

Besides image captioning, Localized Narratives are a natural fit for: (i) image generation: the user can describe which image they want to generate by talking and moving their mouse to indicate the position of objects (demonstration in supp. material, Sec. 1); (ii) image retrieval: the user naturally describes the content of an image they are looking for, in terms of both what and where; (iii) grounded speech recognition: considering the content of an image would allow better speech transcription, e.g. 'plant' and 'planet' are easier to distinguish in the visual than in the voice domain; (iv) voice-driven environment navigation: the user describes where they want to navigate to, using relative spatial language.

To summarize, our paper makes the following contributions: (i) We introduce Localized Narratives, a new form of multimodal image annotations where every word is localized in the image with a mouse trace segment; (ii) We use Localized Narratives to annotate 848,749 images and provide a thorough analysis of the data. (iii) We demonstrate the utility of our data for controlled image captioning.

Dataset	Grounding	Num. captions	Num. images	Words/capt.
COCO Captions [8]	Whole capt. → Whole im.	616,767	123,287	10.5
Conceptual Capt. [46]	Whole capt. → Whole im.	3,334,173	3,334,173	10.3
Stanford Vis. Par. [26]	Whole capt. → Whole im.	19,561	19,561	67.5
ReferIt [24]	Short phrase ⟶ Region	130,525	19,894	3.6
Google Refexp [35]	Short phrase ⟶ Region	104,560	26,711	8.4
Visual Genome [27]	Short phrase ⟶ Region	5,408,689	108,077	5.1
Flickr30k Ent. [40]	Nouns ⟶ Region	158,915	31,783	12.4
Loc. Narr. (Ours)	Each word ⟶ Region	873,107	848,749	36.5

Table 2: **Datasets connecting vision and language via image captioning**, compared with respect to their type of grounding, scale, and caption length. Num. captions is typically higher than num. images because of replication (i.e. several annotators writing a caption for the same image).

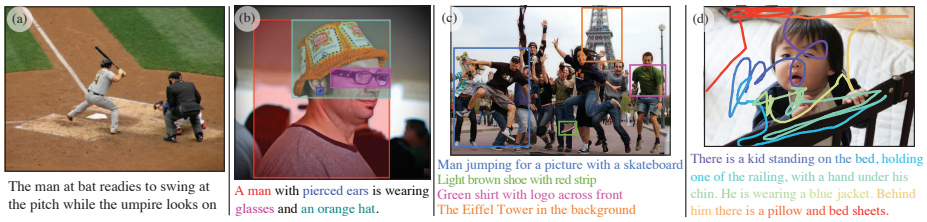


Fig. 2: **Sample annotations** from (a) COCO Captions [8], (b) Flickr30k Entities [40], (c) Visual Genome [27], and (d) Localized Narratives (Ours). For clarity, (b) shows a subset of region descriptions and (d) shows a shorter-than-average Localized Narrative.

2 Related Work

Captioning Datasets. Various annotation efforts connect vision and language via captioning (Tab. 2). We focus on how their captions are grounded, as this is the key differentiating factor of Localized Narratives from these works. As a starting point, classical image captioning [8, 59, 46] and visual paragraph generation [26] simply provide a whole caption for the whole image (Fig. 2(a)). This lack of proper grounding was shown to be problematic [32, 33, 44, 45].

Flickr30k Entities [40] annotated the nouns mentioned in the captions of Flickr30k [59] and drew their bounding box in the image (Fig. 2(b)): the grounding is therefore from nouns to regions (including their attached adjectives, Tab. 2). Visual Genome [27] and related previous efforts [24, 35] provide short phrases describing regions in the images (Fig. 2(c)): grounding is therefore at the phrase level (Tab. 2). While Visual Genome uses these regions as a seed to generate a scene graph, where each node is grounded in the image, the connection between the region descriptions and the scene graph is not explicit.

In Localized Narratives, in contrast, *every word* is grounded to a specific region in the image represented by its trace segment (Fig. 2(d)). This includes all types of words (nouns, verbs, adjectives, prepositions, etc.), in particular valuable spatial-relation markers ("above", "behind", etc.) and relationship indicators ("riding", "holding", etc.). Another disadvantage of Flickr30k Entities

and Visual Genome is that their annotation processes require manually drawing many bounding boxes a posteriori, which is unnatural and time-consuming compared to our simpler and more natural protocol (Sec. 4.1 – Annotation Cost).

SNAG [48] is a proof of concept where annotators describe images using their voice while their gaze is tracked using specialized hardware. This enables inferring the image location they are looking at. As a consequence of the expensive and complicated setup, only 100 images were annotated. In our proposed Localized Narratives, instead, we collect the data using just a mouse, a keyboard, and a microphone as input devices, which are commonly available. This allows us to annotate a much larger set of images (848,749 to date).

In the video domain, ActivityNet-Entities [63] adds visual grounding to the ActivityNet Captions, also in two stages where boxes were drawn a posteriori.

Annotation using Voice. A few recent papers use voice as an input modality for computer vision tasks [11, 49, 48, 18, 17, 16]. The closest work to ours [16] uses voice to simultaneously annotate the class name and the bounding box of an object instance in an image. With Localized Narratives we bring it to the next level by producing richer annotations both in the language and vision domains with long free-form captions associated to synchronized mouse traces.

In the video domain, EPIC-KITCHENS [12] contains videos of daily kitchen activities collected with a head-mounted camera. The actions were annotated with voice, manually transcribed, and time-aligned using YouTube's automatic closed caption alignment tool.

3 Annotation Process

The core idea behind the Localized Narratives annotation protocol is to ask the annotators to describe the contents of the image using their voice while hovering their mouse over the region being described. Both voice and mouse location signals are timestamped, so we know where the annotators are pointing while they are speaking every word.

Figure 3 shows voice (a) and mouse trace data (b), where the color gradient represents temporal synchronization. We summarize how to process this data to produce a Localized Narrative example. First, we apply an Automatic Speech Recognition (ASR) algorithm and get a synchronized, but typically imperfect, transcription (c). After finishing a narration, the annotators transcribe their own recording, which gives us an accurate caption, but without synchronization with the mouse trace (d). Finally, we obtain a correct transcription with timestamps by performing sequence-to-sequence alignment between the manual and automatic transcriptions (e). This time-stamped transcription directly reveals which trace segment corresponds to each word in the caption (f), and completes the creation of a Localized Narrative instance. Below we describe each step in detail.

Annotation Instructions. One of the advantages of Localized Narratives is that it is a natural task for humans to do: speaking and pointing at what we are describing is a common daily-life experience [22, 38]. This makes it easy for annotators to understand the task and perform as expected, while increasing the pool of qualified annotators for the task. The instructions we provide are concise:

Use the mouse to point at the objects in the scene.
Simultaneously, use your voice to describe what you are pointing at.

– Focus on concrete objects (e.g. cow, grass, person, kite, road, sky).
– Do not comment on things you cannot directly see in the image (e.g. feelings that the image evokes, or what might happen in the future).
– Indicate an object by moving your mouse over the whole object, roughly specifying its location and size.
– Say the relationship between two objects while you move the mouse between them, e.g. "a man *is flying* a kite", "a bottle *is on* the table".
– If relevant, also mention attributes of the objects (e.g. *old* car).

Automatic and Manual Transcriptions. We apply an ASR algorithm [14] to obtain an automatic transcription of the spoken caption, which is timestamped but typically contains transcription errors. To fix these errors, we ask the annotators to manually transcribe their own recorded narration. Right after they described an image, the annotation tool plays their own voice recording accompanied by the following instructions:

Type literally what you just said.

– Include filler words if you said them (e.g. "I think", "alright") but not filler sounds (e.g. "um", "uh", "er").
– Feel free to separate the text in multiple sentences and add punctuation.

The manual transcription is accurate but not timestamped, so we cannot associate it with the mouse trace to recover the grounding of each word.

Transcription Alignment. We obtain a correct transcription with timestamps by performing a sequence-to-sequence alignment between the manual and automatic transcriptions (Fig. 3).

Let $\mathbf{a}=\{(a_1,\dots,a_{|\mathbf{a}|}\}$ and $\mathbf{m}=\{m_1,\dots,m_{|\mathbf{m}|}\}$ be the automatic and manual transcriptions of the spoken caption, where a_i and m_j are individual words. a_i is timestamped: let $[t_i^0,t_i^1]$ be the time segment during which a_i was spoken. Our goal is to align $\mathbf{a}$ and $\mathbf{m}$ to transfer the timestamps from the automatically transcribed words a_i to the manually provided m_j.

To do so, we apply Dynamic Time Warping [28] between $\mathbf{a}$ and $\mathbf{m}$. Intuitively, we look for a matching function μ that assigns each word a_i to a word $m_{\mu(i)}$, such that if $i_2>i_1$ then $\mu(i_2)\geq\mu(i_1)$ (it preserves the order of the words). Note that μ assigns each a_i to exactly one m_j, but m_j can match to zero or multiple words in $\mathbf{a}$. We then look for the optimal matching μ^* such that:

$$\mu^* = \arg\min_{\mu} D_\mu(\mathbf{a},\mathbf{m}) \qquad D_\mu(\mathbf{a},\mathbf{m}) = \sum_{i=1}^{|\mathbf{a}|} d(a_i, m_{\mu(i)}) \tag{1}$$

where d is the edit distance between two words, i.e. the number of character inserts, deletes, and replacements required to get from one word to the other. $D_{\mu^*}(\mathbf{a},\mathbf{m})$ provides the optimal matching score (used below to assess quality).

Given μ^*, let the set of matches for m_j be defined as $A_j = \{i \,|\, \mu^*(i)=j\}$. The timestamp $[\bar{t}_j^0,\bar{t}_j^1]$ of word m_j in the manual transcription is the interval

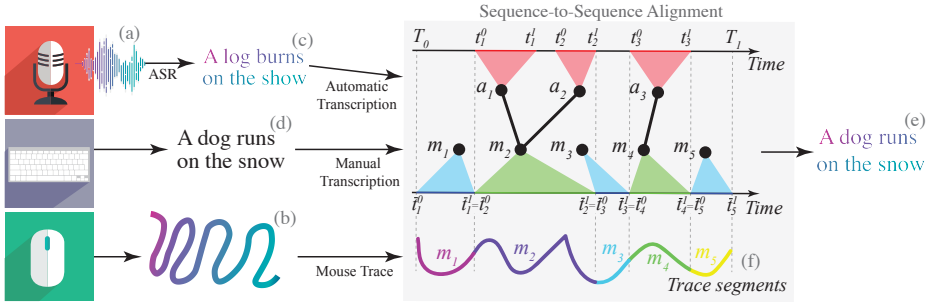


Fig. 3: **Localized Narratives annotation**: We align the automatic transcription (c) to the manual one (d) to transfer the timestamps from the former to the latter, resulting in a transcription that is both accurate and timestamped (e). To do so, we perform a sequence-to-sequence alignment (gray box) between a_i and m_j (black thick lines). The timestamps of matched words m_j are defined as the segment (green) containing the original timestamps (red) of the matched words a_i. Unmatched words m_j get assigned the time segments in between matched neighboring words (blue). These timestamps are transferred to the mouse trace and define the trace segment for each word m_j.

spanned by its matching words (if any) or to the time between neighboring matching words (if none). Formally:

$$\bar{t}_j^0 = \begin{cases} \min\left\{t_i^0 \,|\, i \in A_j\right\} & \text{if } A_j \neq \emptyset, \\ \max\left\{t_i^1 \,|\, i \in A_k \,|\, k < j\right\} & \text{if } \exists\, k < j \text{ s.t. } A_k \neq \emptyset \\ T^0 & \text{otherwise,} \end{cases} \tag{2}$$

$$\bar{t}_j^1 = \begin{cases} \max\left\{t_i^1 \,|\, i \in A_j\right\} & \text{if } A_j \neq \emptyset, \\ \min\left\{t_i^0 \,|\, i \in A_k \,|\, k > j\right\} & \text{if } \exists\, k > j \text{ s.t. } A_k \neq \emptyset \\ T^1 & \text{otherwise,} \end{cases}$$

where T^0 is the first time the mouse pointer *enters* the image and T^1 is the last time it *leaves* it. Finally, we define the *trace segment* associated with a word m_j as the segment of the mouse trace spanned by the time interval $[\bar{t}_j^0, \bar{t}_j^1]$ (Fig. 3).

Automatic quality control. To ensure high-quality annotations, we devise an automatic quality control mechanism by leveraging the fact that we have a double source of voice transcriptions: the manual one given by the annotators (**m**) and the automatic one given by the ASR system (**a**, Fig. 3). We take their distance $D_{\mu^*}(\mathbf{a}, \mathbf{m})$ in the optimal alignment μ^* as a quality control metric (Eq. (1)). A high value of D_{μ^*} indicates large discrepancy between the two transcriptions, which could be caused by the annotator having wrongly transcribed the text, or due to the ASR failing to recognize the annotators' spoken words. In contrast, a low value of D_{μ^*} indicates that the transcription is corroborated by two sources. In practice, we manually analyzed a large number of annotations at different values of D_{μ^*} and choose a specific threshold below which essentially all transcriptions were correct. We discarded all annotations above this threshold.

In addition to this automatic quality control, we also evaluate the quality of the annotations in terms of semantic accuracy, visual grounding accuracy, and quality of manual voice transcription (Sec. 4.2).

4 Dataset Collection, Quality, and Statistics

4.1 Dataset collection

Image Sources and Scale. We annotated a total of 848,749 images with Localized Narratives over 4 datasets: (i) COCO [31, 8] (train and validation, 123k images); (ii) Flickr30k [59] (train, validation, and test, 32k); (iii) ADE20K [62] (train and validation, 20k); (iv) Open Images (full validation and test, 167k, and part of train, 504k). For Open Images, to enable cross-modal applications, we selected images for which object segmentations [5], bounding boxes or visual relationships [29] are already available. We annotated 5,000 randomly selected COCO images with replication 5 (i.e. 5 different annotators annotated each image). Beyond this, we prioritized having a larger set covered, so the rest of images were annotated with replication 1. All analyses in the remainder of this section are done on the full set of 849k images, unless otherwise specified.

Annotation Cost. Annotating one image with Localized Narratives takes 144.7 seconds on average. We consider this a relatively low cost given the amount of information harvested, and it allows data collection at scale. Manual transcription takes up the majority of the time (104.3 sec., 72%), while the narration step only takes 40.4 seconds (28%). In the future, when ASR systems improve further, manual transcription could be skipped and Localized Narratives could become even faster thanks to our core idea of using speech.

To put our timings into perspective, we can roughly compare to Flickr30k Entities [40], which is the only work we are aware of that reports annotation times. They first manually identified which words constitute entities, which took 235 seconds per image. In a second stage, annotators drew bounding boxes for these selected entities, taking 408 seconds (8.7 entities per image on average). This yields a total of 643 seconds per image, without counting the time to write the actual captions (not reported). This is 4.4× slower than the total annotation cost of our method, which includes the grounding of 10.8 nouns per image and the writing of the caption. The Visual Genome [27] dataset was also annotated by a complex multi-stage pipeline, also involving drawing a bounding box for each phrase describing a region in the image.

4.2 Dataset Quality

To ensure high quality, Localized Narratives was made by 156 professional annotators working full time on this project. Annotator managers did frequent manual inspections to keep quality consistently high. In addition, we used an automatic quality control mechanism to ensure that the spoken and written transcriptions match (Sec. 3 – Automatic quality control). In practice, we placed a high quality bar, which resulted in discarding 23.5% of all annotations. Below we analyze the quality of the annotations that remained after this automatic discarding step (all dataset statistics reported in this paper are after this step too).

Semantic and Transcription Accuracy. In this section we quantify (i) how well the noun phrases and verbs in the caption correctly represent the objects in

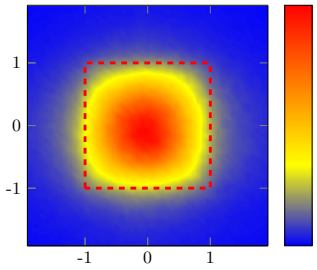


Fig. 4: **Mouse trace segment locations** on COCO with respect to the closest box of the relevant class (- - -).

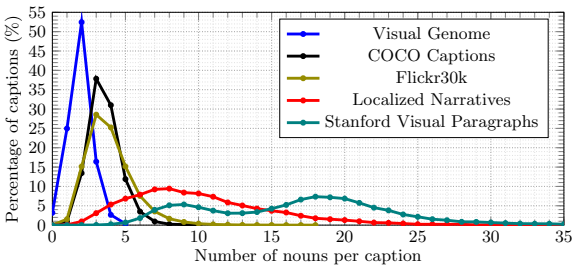


Fig. 5: **Distribution of number of nouns per caption**. As in Table 3, these counts are per individual caption.

the image (Semantic accuracy) and (ii) how well the manually transcribed caption matches the voice recording (Transcription accuracy). We manually check every word in 100 randomly selected Localized Narratives on COCO and log each of these two types of errors. This was done carefully by experts (authors of this paper), not by the annotators themselves (hence an independent source).

In terms of semantic accuracy, we check every noun and verb in the 100 captions and assess whether that object or action is indeed present in the image. We allow generality up to a base class name (e.g. we count either "dog" or "Chihuahua" as correct for a Chihuahua in the image) and we strictly enforce correctness (e.g. we count "skating" as incorrect when the correct term is "snowboarding" or "bottle" in the case of a "jar"). Under these criteria, semantic accuracy is very high: 98.0% of the 1,582 nouns and verbs are accurate.

In terms of transcription accuracy, we listen to the voice recordings and compare them to the manual transcriptions. We count every instance of (i) a missing word in the transcription, (ii) an extra word in the transcription, and (iii) a word with typographical errors. We normalize these by the total number of words in the 100 captions (4,059). This results in 3.3% for type (i), 2.2% for (ii), and 1.1% for (iii), showing transcription accuracy is high.

Localization Accuracy. To analyze how well the mouse traces match the location of actual objects in the image, we extract all instances of any of the 80 COCO object classes in our captions (exact string matching, 600 classes in the case of Open Images). We recover 146,723 instances on COCO and 374,357 on Open Images train. We then associate each mouse trace segment to the closest ground-truth box of its corresponding class. Figure 4 displays the 2D histogram of the positions of all trace segment points with respect to the closest box (- - -), normalized by box size for COCO. We observe that most of the trace points are within the correct bounding box (the figure for Open Images is near-identical, see supp. material Sec. 3).

We attribute the trace points that fall outside the box to two different effects. First, circling around the objects is commonly used by annotators (Fig. 1 and Fig. 6). This causes the mouse traces to be close to the box, but not inside it. Second, some annotators sometimes start moving the mouse before they describe

the object, or vice versa. We see both cases as a research opportunity to better understand the connection between vision and language.

4.3 Dataset Statistics

Richness. The mean length of the captions we produced is 36.5 words (Tab. 2), substantially longer than all previous datasets, except Stanford Visual Paragraphs [26] (e.g. 3.5× longer than the individual COCO Captions [8]). Both Localized Narratives and Stanford Visual Paragraphs describe an image with a whole paragraph, as opposed to one sentence [8, 46, 24, 35, 27, 59]. However, Localized Narratives additionally provide dense visual grounding via a mouse trace segment for each word, and has annotations for 40× more images than Stanford Visual Paragraphs (Tab.2).

We also compare in terms of the average number of nouns, pronouns, adjectives, verbs, and adpositions (prepositions and postpositions, Tab. 3). We determined this using the spaCy [19] part-of-speech tagger. Localized Narratives has a higher occurrence per caption for each of these categories compared to most previous datasets, which indicates that our annotations provide richer use of natural language in connection to the images they describe.

Diversity. To illustrate the diversity of our captions, we plot the distribution of the number of nouns per caption, and compare it to the distributions obtained over previous datasets (Fig. 5). We observe that the range of number of nouns is significantly higher in Localized Narratives than in COCO Captions, Flickr30k, Visual Genome, and comparable to Stanford Visual Paragraphs. This poses an additional challenge for captioning methods: automatically adapting the length of the descriptions to each image, as a function of the richness of its content. Beyond nouns, Localized Narratives provide visual grounding for every word (verbs, prepositions, etc.). This is especially interesting for relationship words, e.g. "woman holding ballon" (Fig. 1) or "with a hand under his chin" (Fig. 2(d)). This opens the door to a new venue of research: understanding how humans naturally ground visual relationships.

Diversity in Localized Narratives is present not only in the language modality, but also in the visual modality, such as the different ways to indicate the spatial location of objects in an image. In contrast to previous works, where the

Dataset	Words	Nouns	Pronouns	Adjectives	Adpositions	Verbs
Visual Genome [27]	5.1	1.9	0.0	0.6	0.7	0.3
COCO Captions [8]	10.5	3.6	0.2	0.8	1.7	0.9
Flickr30k [59]	12.4	3.9	0.3	1.1	1.8	1.4
Localized Narratives	36.5	10.8	3.6	1.6	4.7	4.2
Stanford Visual Paragraphs [26]	61.9	17.0	2.7	6.6	8.0	4.1

Table 3: **Richness of individual captions** of Localized Narratives versus previous works. Please note that since COCO Captions and Flickr30K have replication 5 (and Visual Genome also has a high replication), counts *per image* would be higher in these datasets. However, many of them would be duplicates. We want to highlight the richness of captions as units and thus we show word counts averaged over *individual captions*.

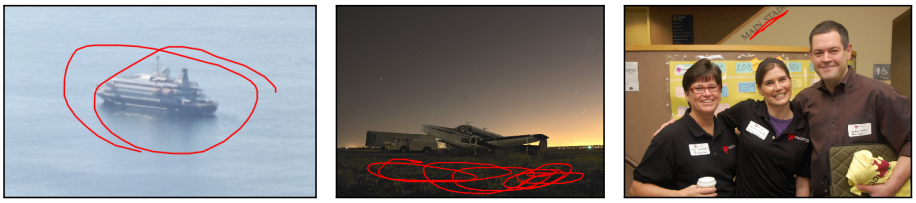


Fig. 6: **Examples of mouse trace segments and their corresponding word(s) in the caption** with different pointing styles: circling, scribbling, and underlining.

grounding is in the form of a bounding box, our instructions lets the annotator hover the mouse over the object in any way they feel natural. This leads to diverse styles of creating trace segments (Fig. 6): circling around an object (sometimes without even intersecting it), scribbling over it, underlining in case of text, etc. This diversity also presents another challenge: detect and adapt to different trace styles in order to make full use of them.

5 Controlled Image Captioning

We now showcase how localized narratives can be used for controlled image captioning. Controlled captioning was first proposed in [10] and enables a user to specify which parts of the image they want to be described, and in which order. In [10] the user input was in the form of user-provided bounding boxes. In this paper we enable controllability through a mouse trace, which provides a more intuitive and efficient user interface. One especially useful application for controlled captioning is assistive technology for people with imperfect vision [6, 54, 61], who could utilize the mouse to express their preferences in terms of how the image description should be presented.

Task definition. Given both an image and a mouse trace, the goal is to produce an image caption which matches the mouse trace, i.e. it describes the image regions covered by the trace, and in the order of the trace. This task is illustrated by several qualitative examples of our controlled captioning system in Fig. 7. In both the image of the skiers and the factory, the caption correctly matches the given mouse trace: it describes the objects which were indicated by the user, in the order which the user wanted.

Method. We start from a state-of-the-art, transformer-based encoder-decoder image captioning model [3, 7]. This captioning model consumes Faster-RCNN features [43] of the top 16 highest scored object proposals in the image. The Faster-RCNN module is pre-trained on Visual Genome [27] (excluding its intersection with COCO). The model uses these features to predict an image caption based on an attention model, inspired by the Bottom-Up Top-Down approach of [3]. This model is state of the art for standard image captioning, i.e. it produces captions given images alone as input (Fig. 8(a)).

We modify this model to also input the mouse trace, resulting in a model that consumes four types of features both at training time and test time: (i)

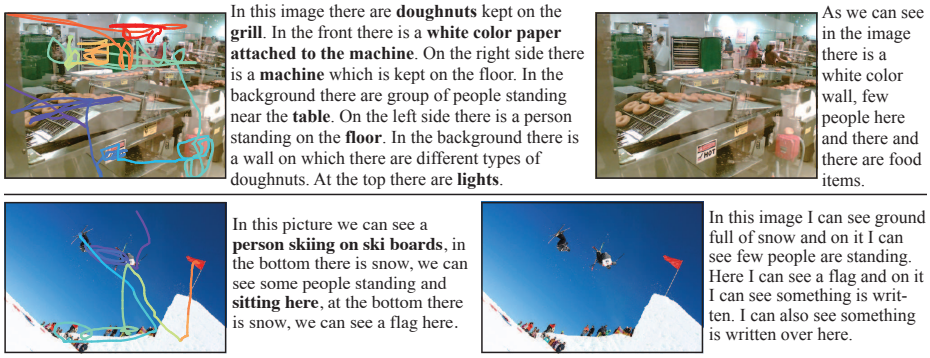
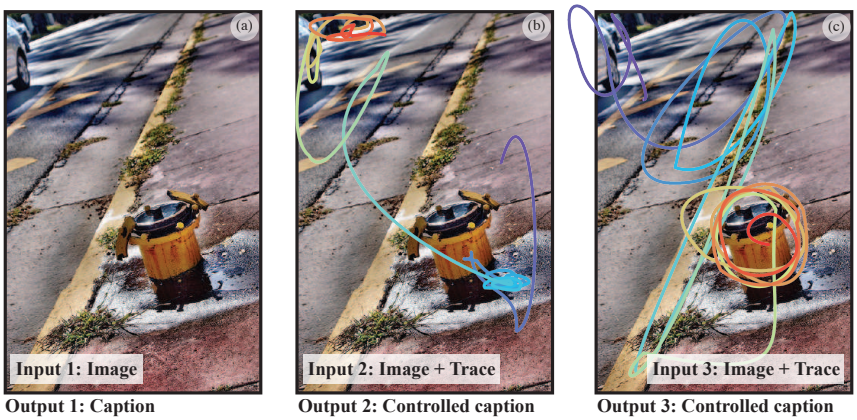


Fig. 7: **Qualitative results for controlled image captioning.** Gradient indicates time. Captioning controlled by mouse traces (left) and without traces (right). The latter misses important objects: e.g. skiers in the sky, doughnuts – all in bold.

Faster R-CNN features of the automatically-detected top object proposals, representing their semantic information; (ii) the coordinate and size features of these proposals, representing the location of the detected objects. (iii) the total time duration of the mouse trace, capturing information about the expected length of the full image description. (iv) the position of the mouse trace as it moves over the image, representing the visual grounding. To create this representation, we first divide the mouse trace evenly into pseudo-segments based on the prior median word duration (0.4 sec over the whole training set). We then represent each pseudo-segment by its encapsulating bounding box, resulting in a set of features which take the same form as (ii). This new model takes an image *plus* a mouse trace as input and produces the caption that the user is interested in. More technical details in the supp. material, Sec. 4.

Evaluation. Our first metric is the standard ROUGE-L [30]. This metric determines the longest common subsequence (LCS) of words between the predicted caption and the reference caption, and calculates the F1-score (harmonic mean over precision and recall of words in the LCS). This means ROUGE-L explicitly measures word order. We also measure the F1 score of ROUGE-1, which we term ROUGE-1-F1. This measures the F1-score w.r.t. co-occurring words. Hence ROUGE-1-F1 is the orderless counterpart of ROUGE-L and enables us to separate the effects of caption completeness (the image parts which the user wanted to be described) and word order (the order in which the user wanted the image to be described). For completeness we also report other standard captioning metrics: BLEU-1, BLEU-4 [39], CIDEr-D [50], and SPICE [2]. For all measures, a higher number reflects a better agreement between the caption produced by the model and the ground-truth caption written by the annotator.

We observe that in standard image captioning tasks there typically are multiple reference captions to compare to [8, 50, 59], since that task is ambiguous: it is unclear what image parts should be described and in which order. In contrast, our controlled image captioning task takes away both types of ambiguity, resulting in a much better defined task. As such, in this evaluation we compare



Output 1: Caption
There is a fire hydrant and this is road.

Output 2: Controlled caption
In this image, we can see a platform, there is a yellow color pot on it. Left side, there is a road, car, few trees we can see on the right side.

Output 3: Controlled caption
This image consists of a car parked on the road. To the top left, there is a car. To the right, there is a footpath on which a fire hydrant is kept.

Fig. 8: **Qualitative results for controlled image captioning.** Standard (a) versus controlled captioning (b) and (c). (a) misses important objects such as the car or the footpath. In (b) and (c) the controlled output captions adapt to the order of the objects defined by the trace. Gradient indicates time. More in the supp. mat. Sec. 2.

to a single reference only: given an image plus a human-provided mouse trace, we use its corresponding human-provided caption as reference.

Results. We perform experiments on the Localized Narratives collected on COCO images, using the standard 2017 training and validation splits. To get a feeling of what a trace can add to image captioning, we first discuss the qualitative examples in Fig. 7 and 8. First of all, the trace focuses the model attention on specific parts of the image, leading it to mention objects which would otherwise be missed: In the top-left image of Fig. 7, the trace focuses attention on the skiers, which are identified as such (in contrast to the top-right). Similarly, the top-left and right of Fig. 7, using the trace results in focusing on specific details which leads to more complete and more fine-grained descriptions (e.g. doughnuts, grill, machine, lights). Finally in Fig. 8a, the standard captioning model misses the car since it is not prominent in the image. In Fig 8b and c instead, the augmented model sees both traces going over the car and produces a caption including it. In this same figure, we can also see that different traces lead to different captions. These results suggests that conditioning on the trace helps with covering the image more completely and highlighting specific objects within it. At the same time, we can see in all examples that the trace order maps nicely to the word order in the caption, which is the order the user wanted.

Table 4 shows quantitative results. Compared to standard captioning [3, 7], all metrics improve significantly when doing controlled captioning using the mouse trace. BLEU-4 and CIDEr-D are particularly affected and improve by more than 3×. ROUGE-1-F1 increased from 0.479 for standard captioning to 0.607 for controlled captioning using the full mouse trace. Since ROUGE-1-F1 ignores word order, this increase is due to the *completeness* of the caption only: it indicates

Method	features	ROUGE-L	ROUGE-1-F1	BLEU-1	BLEU-4	CIDEr-D	SPICE
Standard captioning [3, 7]	**i**	**0.317**	**0.479**	**0.322**	**0.081**	**0.293**	**0.257**
+ proposal locations	i+ii	0.318	0.482	0.323	0.082	0.295	0.257
+ mouse trace duration	i+ii+iii	0.334	0.493	0.372	0.097	0.373	0.265
Controlled captioning	**i+ii+iii+iv**	**0.483**	**0.607**	**0.522**	**0.246**	**1.065**	**0.365**

Table 4: **Controlled image captioning results on the COCO validation set**, versus standard (non-controlled) captioning, and two ablations.

that using the mouse trace enables the system to better describe those parts of the image which were indicated by the user.

Switching from ROUGE-1-F1 to ROUGE-L imposes a word order. The standard captioning model yields a ROUGE-L of 0.317, a drop of 34% compared to ROUGE-1-F1. Since standard captioning does not input any particular order within the image (but does use a linguistically plausible ordering), this drop can be seen as a baseline for not having information on the order in which the image should be described. When using the mouse trace, the controlled captioning model yields a ROUGE-L of 0.483, which is a much smaller drop of 20%. This demonstrates quantitatively that our controlled captioning model successfully exploits the input trace to determine the order in which the user wanted the image to be described. Overall, the controlled captioning model outperforms the standard captioning model by 0.166 ROUGE-L on this task (0.483 vs 0.317).

Ablations. We perform two ablations to verify whether most improvements indeed come from the mouse trace itself, as opposed to the other features we added. Starting from standard captioning, we add the locations of the object proposals from which the model extracts visual features (Tab. 4, "+ proposal locations", feature (ii)). This has negligible effects on performance, suggesting that this model does not benefit from knowing where in the image its appearance features (i) came from. Next, we add the trace time duration (Tab. 4, "+ mouse trace duration"). This gives an indication of how long the caption the user wants should be. This brings minor improvements only. Hence, most improvements come when using the full mouse trace, demonstrating that most information comes from the location and order of the mouse trace (Tab. 4, controlled captioning).

Summary. To summarize, we demonstrated that using the mouse trace leads to large improvements when compared to a standard captioning model, for the task of controlled captioning. Importantly, we do not claim the resulting captions are better in absolute terms. Instead, they are better fitting what the user wanted, in terms of which parts of the image are described and in which order.

6 Conclusions

This paper introduces Localized Narratives, an efficient way to collect image captions in which every single word is visually grounded by a mouse trace. We annotated 849k images with Localized Narratives. Our analysis shows that our data is rich and provides accurate grounding. We demonstrate the utility of our data through controlled image captioning using the mouse trace.

References

1. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al.: Deep speech 2: End-to-end speech recognition in English and Mandarin. In: ICML (2016)
2. Anderson, P., Fernando, B., Johnson, M., Gould, S.: SPICE: semantic propositional image caption evaluation. In: ECCV (2016)
3. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018)
4. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: visual question answering. In: ICCV (2015)
5. Benenson, R., Popov, S., Ferrari, V.: Large-scale interactive object segmentation with human annotators. In: CVPR (2019)
6. Bigham, J.P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R.C., Miller, R., Tatarowicz, A., White, B., White, S., Yeh, T.: VizWiz: nearly real-time answers to visual questions. In: Proceedings of the 23nd annual ACM symposium on User interface software and technology (2010)
7. Changpinyo, S., Pang, B., Sharma, P., Soricut, R.: Decoupled box proposal and featurization with ultrafine-grained semantic labels improve image captioning and visual question answering. In: EMNLP-IJCNLP (2019)
8. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft COCO captions: Data collection and evaluation server. arXiv (2015)
9. Cirik, V., Morency, L.P., Berg-Kirkpatrick, T.: Visual referring expression recognition: What do systems actually learn? In: NAACL (2018)
10. Cornia, M., Baraldi, L., Cucchiara, R.: Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In: CVPR (2019)
11. Dai, D.: Towards Cost-Effective and Performance-Aware Vision Algorithms. Ph.D. thesis, ETH Zurich (2016)
12. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: The EPIC-KITCHENS dataset: Collection, challenges and baselines. IEEE Trans. on PAMI (2020)
13. Dogan, P., Sigal, L., Gross, M.: Neural sequential phrase grounding (seqground). In: CVPR (2019)
14. Google cloud speech-to-text API. `https://cloud.google.com/speech-to-text/`
15. Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: ICASSP (2013)
16. Gygli, M., Ferrari, V.: Efficient object annotation via speaking and pointing. IJCV (2019)
17. Gygli, M., Ferrari, V.: Fast Object Class Labelling via Speech. In: CVPR (2019)
18. Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A., Glass, J.: Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input. In: ECCV (2018)
19. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017), `spacy.io`
20. Hudson, D.A., Manning, C.D.: GQA: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR (2019)
21. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: CVPR (2016)

22. Kahneman, D.: Attention and effort. Citeseer (1973)
23. Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., Oord, A.v.d., Dieleman, S., Kavukcuoglu, K.: Efficient neural audio synthesis. In: ICML (2018)
24. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: EMNLP (2014)
25. Kim, D.J., Choi, J., Oh, T.H., Kweon, I.S.: Dense relational captioning: Triple-stream networks for relationship-based captioning. In: CVPR (2019)
26. Krause, J., Johnson, J., Krishna, R., Fei-Fei, L.: A hierarchical approach for generating descriptive image paragraphs. In: CVPR (2017)
27. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV **123**(1), 32–73 (2017)
28. Kruskal, J.B., Liberman, M.: The symmetric time-warping problem: from continuous to discrete. In: Time Warps, String Edits, and Macromolecules - The Theory and Practice of Sequence Comparison, chap. 4. CSLI Publications (1999)
29. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., Ferrari, V.: The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. arXiv preprint arXiv:1811.00982 (2018)
30. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out (2004)
31. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft COCO: Common objects in context. In: ECCV (2014)
32. Liu, C., Mao, J., Sha, F., Yuille, A.: Attention correctness in neural image captioning. In: AAAI (2017)
33. Lu, J., Yang, J., Batra, D., Parikh, D.: Neural baby talk. In: CVPR (2018)
34. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A neural-based approach to answering questions about images. In: ICCV (2015)
35. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: CVPR (2016)
36. Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., Courville, A., Bengio, Y.: Samplernn: An unconditional end-to-end neural audio generation model. In: ICLR (2017)
37. Oord, A.v.d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: Wavenet: A generative model for raw audio. arXiv **1609.03499** (2016)
38. Oviatt, S.: Multimodal interfaces. The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications (2003)
39. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: ACL (2002)
40. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. IJCV **123**(1), 74–93 (2017)
41. Ravanelli, M., Parcollet, T., Bengio, Y.: The pytorch-kaldi speech recognition toolkit. In: ICASSP (2019)
42. Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. In: NeurIPS. pp. 217–225 (2016)

43. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NeurIPS (2015)
44. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: EMNLP (2018)
45. Selvaraju, R.R., Lee, S., Shen, Y., Jin, H., Ghosh, S., Heck, L., Batra, D., Parikh, D.: Taking a HINT: Leveraging explanations to make vision and language models more grounded. In: ICCV (2019)
46. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018)
47. Tan, F., Feng, S., Ordonez, V.: Text2scene: Generating compositional scenes from textual descriptions. In: CVPR (2019)
48. Vaidyanathan, P., Prud, E., Pelz, J.B., Alm, C.O.: SNAG : Spoken Narratives and Gaze Dataset. In: ACL (2018)
49. Vasudevan, A.B., Dai, D., Van Gool, L.: Object Referring in Visual Scene with Spoken Language. In: CVPR (2017)
50. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: CIDEr: Consensus-based image description evaluation. In: CVPR (2015)
51. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR (2015)
52. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE Trans. on PAMI **39**(4), 652–663 (2016)
53. Website: Localized Narratives Data and Visualization. `https://google.github.io/localized-narratives` (2020)
54. Wu, S., Wieland, J., Farivar, O., Schiller, J.: Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In: Conference on Computer Supported Cooperative Work and Social Computing (2017)
55. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML (2015)
56. Yan, S., Yang, H., Robertson, N.: ParaCNN: Visual paragraph generation via adversarial twin contextual CNNs. arXiv (2020)
57. Yin, G., Liu, B., Sheng, L., Yu, N., Wang, X., Shao, J.: Semantics disentangling for text-to-image generation. In: CVPR (2019)
58. Yin, G., Sheng, L., Liu, B., Yu, N., Wang, X., Shao, J.: Context and attribute grounded dense captioning. In: CVPR (2019)
59. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL **2**, 67–78 (2014)
60. Yu, J., Li, J., Yu, Z., Huang, Q.: Multimodal transformer with multi-view visual representation for image captioning. arXiv **1905.07841** (2019)
61. Zhao, Y., Wu, S., Reynolds, L., Azenkot, S.: The effect of computer-generated descriptions on photo-sharing experiences of people with visual impairments. ACM on Human-Computer Interaction (2017)
62. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ADE20K dataset. IJCV **127**(3), 302–321 (2019)
63. Zhou, L., Kalantidis, Y., Chen, X., Corso, J.J., Rohrbach, M.: Grounded video description. In: CVPR (2019)
64. Ziegler, Z.M., Melas-Kyriazi, L., Gehrmann, S., M. Rush, A.: Encoder-agnostic adaptation for conditional language generation. arXiv (2019)